

# 3HANDS Dataset: Learning from Humans for Generating Naturalistic Handovers with Supernumerary Robotic Limbs

**Conference Paper****Author(s):**

Saberpour Abadian, Artin; Liao, Yi-Chi; Otaran, Ata; Dabral, Rishabh; Muehlhaus, Marie; Theobalt, Christian; Schmitz, Martin; Steimle, Jürgen

**Publication date:**

2025

**Permanent link:**

<https://doi.org/https://doi.org/10.3929/ethz-b-000738519>

**Rights / license:**

[Creative Commons Attribution-NonCommercial 4.0 International](#)

**Originally published in:**

<https://doi.org/10.1145/3706598.3713306>



# 3HANDS Dataset: Learning from Humans for Generating Naturalistic Handovers with Supernumerary Robotic Limbs

Artin Saberpour Abadian  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
saberpour@cs.uni-saarland.de

Yi-Chi Liao  
ETH Zürich  
Zürich, Switzerland  
yichi.liao@inf.ethz.ch

Ata Otaran  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
otaran@cs.uni-saarland.de

Rishabh Dabral  
Max Planck Institute for  
Informatics  
Saarbrücken, Germany  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
rdabral@mpi-inf.mpg.de

Marie Muehlhaus  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
muehlhaus@cs.uni-saarland.de

Christian Theobalt  
Max Planck Institute for  
Informatics  
Saarbrücken, Germany  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
theobalt@mpi-inf.mpg.de

Martin Schmitz  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
mschmitz@cs.uni-saarland.de

Jürgen Steimle  
Saarland University  
Saarland Informatics Campus  
Saarbrücken, Germany  
steimle@cs.uni-saarland.de

## Abstract

Supernumerary robotic limbs are robotic structures integrated closely with the user's body, which augment human physical capabilities and necessitate seamless, naturalistic human-machine interaction. For effective assistance in physical tasks, enabling SRLs to hand over objects to humans is crucial. Yet, designing heuristic-based policies for robots is time-consuming, difficult to generalize across tasks, and results in less human-like motion. When trained with proper datasets, generative models are powerful alternatives for creating naturalistic handover motions. We introduce 3HANDS, a novel dataset of object handover interactions between a participant performing a daily activity and another participant enacting a hip-mounted SRL in a naturalistic manner. 3HANDS captures the unique characteristics of SRL interactions: operating in intimate personal space with asymmetric object origins, implicit motion synchronization, and the user's engagement in a primary task during the handover. To demonstrate the effectiveness of our dataset, we present three models: one that generates naturalistic handover trajectories, another that determines the appropriate handover endpoints, and a third that predicts the moment to initiate a handover. In a user study (N=10), we compare the handover

interaction performed with our method compared to a baseline. The findings show that our method was perceived as significantly more natural, less physically demanding, and more comfortable.

## CCS Concepts

• **Human-centered computing** → **User centered design; Gestural input**; • **Computing methodologies** → **Robotic planning; Spatial and physical reasoning**.

## Keywords

supernumerary robotic limb, wearable robotic arm, third arm, handover, dataset, motion synthesis, generative model, data-driven control in robotics

## ACM Reference Format:

Artin Saberpour Abadian, Yi-Chi Liao, Ata Otaran, Rishabh Dabral, Marie Muehlhaus, Christian Theobalt, Martin Schmitz, and Jürgen Steimle. 2025. 3HANDS Dataset: Learning from Humans for Generating Naturalistic Handovers with Supernumerary Robotic Limbs. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3706598.3713306>

## 1 Introduction

Supernumerary robotic limbs (SRLs) hold great promise in supporting humans in diverse activities by seamlessly integrating human bodies with assistive motion. Frequently investigated applications include physical activities where an additional hand is needed [16], physical assistance for the elderly [59], augmenting humans with "superpowers" [78], or assistance for strenuous physical tasks [29, 50, 62, 76].



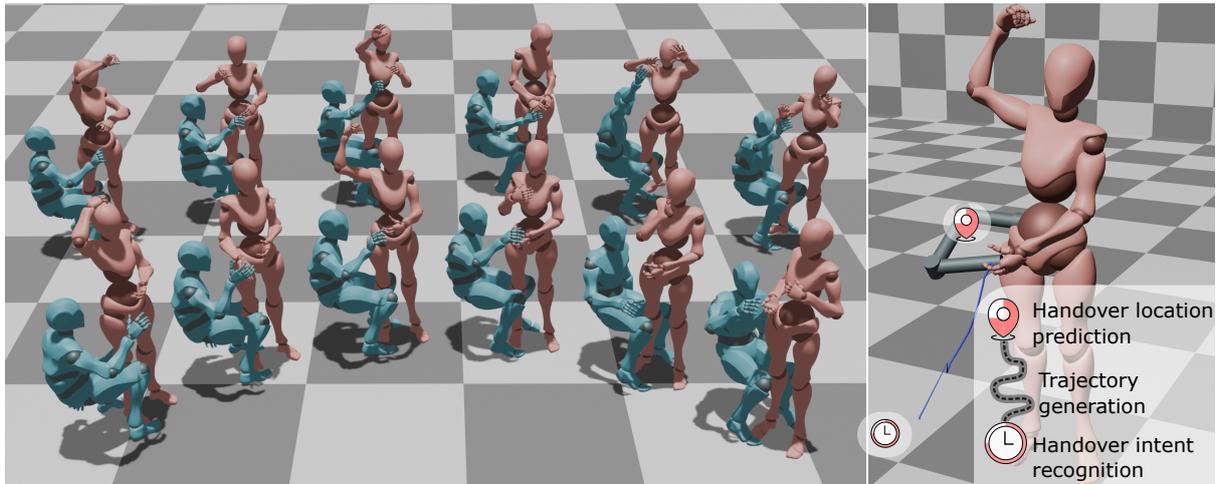
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3713306>



**Figure 1: The 3HANDS dataset comprises an extensive collection of human motion data of asymmetric object handovers between users and a human-enacted third arm, which assists an ongoing activity by handing in or taking away objects at an intimate distance to the user. It contains recordings of 946 interactions captured with 12 participant pairings while performing 12 daily activities. The dataset comprises rigged skeleton data of full body (69 joints) and hands (21 joints). We demonstrate the dataset’s utility to train state-of-the-art machine learning models for three essential steps in the handover activity: generating naturalistic handover trajectories, predicting the location of the handover, and identifying the intent to initialize a handover.**

Prior works also found SRL’s wide-range applications in the Human-Computer Interaction (HCI) field, such as holding support, offering additional control [44], providing rich haptic feedback [35], and complex human-robot collaboration [77]. In all these cases, handing over objects between human hands and the robotic limb is a frequent activity.

Given that SRLs are human-machine interfaces that operate within personal and often even intimate distance to the user [21, 41], their motion control demands are particularly stringent to ensure interactions that are predictable, safe, and effective. One might intuitively attempt to design handover motions for SRLs based on heuristics; however, this approach is a tedious control and programming task [52]. It also can easily fail to account for the subtleties of natural human interaction, including naturalistic patterns of motion kinematics acceptable in intimate personal space, effective inter-hand motion coordination, and subtle cues that convey handover intention. Modern generative machine learning techniques offer a promising alternative, allowing us to develop data-driven models capable of generating natural and safe motions for user interaction [60, 61]. Achieving this goal hinges on the availability of appropriate datasets for training these models.

Several handover datasets currently exist. For instance, the H2O dataset [89] captures the hand postures at a short distance of both the giver and the receiver in front of each other, whereas the HOH dataset captures whole-body motions as two users sit face-to-face [83]. Others captured bimanual handover motions [36]. While these datasets provide valuable insights into handover motions, they do not account for the differences in interactions and motion kinematics arising from operating in the user’s personal space.

Prior datasets studied face-to-face handovers with symmetric roles initiated based on clear temporal cues and without another primary activity. In contrast, the handover with an SRL is characterized by (1) asymmetric spatial configuration centered on the user’s body, (2) asymmetric roles of SRL (assistant) and user (master), (3) ongoing primary activity of the user, potentially influencing timing and location of handover, and (4) implicit initiation of handover based on the user’s implicit postural cues. These substantial differences demand a novel dataset that addresses these critical factors, which can then be used to train models for generating naturalistic SRL motions.

This paper contributes 3HANDS, a human-human object handover dataset specifically developed to help design the interactive behavior of hip-mounted supernumerary robotic limbs. It captures interacting pairs of humans (see Figure 1). One person is performing 12 different daily activities, sampled from activities at the torso and above the shoulders, close or far from the body, and with a small to large range of motion, aiming to cover a broad spectrum of human motion dynamics. Examples range from shampooing hair and hammering to painting the wall and cleaning a window. Concurrently, a second person acted as the SRL and was instructed to hand over and take back a spherical object to the first person in a natural manner, taking a position and posture representative of a hip-mounted SRL. We opted for a hip-mounted configuration, as it is a common SRL mounting location (e.g., [59, 64, 80]), more stable for mechanical movements [80], minimizes interference with the user’s natural arm workspace, and enhances safety by distancing the SRL from sensitive areas (like the head and face). We captured 946 interactions performed by 12 unique pairings of participants. The dataset was captured using

a markerless motion-capture setup, with 41 synchronized 2K camera views. Markerless capture is considerably less invasive than the marker-based setups used in many prior studies and ensured that participants could move freely and naturally to perform the desired tasks without restraining their motions. The dataset contains detailed skeleton data of 69 joints with 107 Degrees of Freedom (DoF), including the detailed capture of hand articulation with 21 joints per hand. We recorded the participants' rigged 3D skeletons, hand poses, time-synchronized textual transcriptions of verbal utterances, information on whether a handover is occurring and the transcription of the verbal communication between participants with the time stamps for each interaction. Our dataset has four key characteristics that distinguish it from prior datasets and make it particularly proper for training SRLs: Contrasting with prior datasets that captured face-to-face handover with a clear handover temporal cue [83, 89], our dataset has the following features that made it particularly proper for training SRLs: (1) Instead of face-to-face motions, handovers occur in an *asymmetric spatial configuration* and *in intimate distance to the user*, where the objects are asymmetrically delivered from the sides of the primary user. (2) Participants take on *asymmetric roles*: primary user vs. robotic assistant. (3) The primary user performs an ongoing *primary activity*. (4) We opted against a specific cue after which both participants should initiate the motion immediately; instead, we precisely capture how participants *implicitly coordinate* the start of a handover. This rich multi-modal dataset offers a valuable resource for the HCI community to investigate the complex interplay between human motion and verbal communication during handovers, ultimately informing the design of more intuitive and user-friendly SRL interfaces in particular and of human-robot interfaces in general. We share the dataset with the community<sup>1</sup>.

To further demonstrate practical applications of our dataset in training models for interaction with SRLs, we trained three distinct models using conditional variational autoencoder (CVAE) [69] and neural network architectures. Each of them addresses one essential step in the handover activity. First, we developed a trajectory generation model capable of generating naturalistic handover motions for SRLs in response to the primary user's actions. Second, we contribute a model to anticipate the desired location where the handover will most likely occur for a given posture. Finally, we show that our dataset facilitates the training of a model that accurately predicts when the SRL should initiate a handover solely based on implicit postural cues of the primary user. We detail on the data processing, models and experimental results. The performance metrics achieved with our dataset confirm its quality and show its potential to both, advance the field of SRLs and deepen the understanding of handover activities in close personal space. Furthermore, we conducted a user study examining the subjective perceived quality of the generated handover motions (for measures such as perceived naturalness, smoothness, and predictability) compared to a baseline method in a virtual reality environment. The results of the study indicated that

our models trained with 3HANDS result in more natural and smooth motions that are less physically demanding and more comfortable. We hope our dataset and experimental results will provide a valuable resource for future studies and applications.

In summary, this paper contributes the following:

- We introduce the 3HANDS dataset, an extensive collection of motion patterns originating from two persons engaging in an object handover. It captures 946 asymmetric handover motions in scenarios where the user is performing a primary activity. It offers a rich set of motion data, comprising rigged 3D skeletons and hand poses, transcriptions of verbal utterances, and information on whether a handover is occurring.
- We illustrate the effectiveness of using the 3HANDS dataset to train models for handover interactions with supernumerary robotic limbs. These a) generate naturalistic handover motion trajectories, b) predict the location of a handover, and c) accurately predict when to initiate a handover.
- In a controlled user study in a virtual reality environment, we verify the naturalness of the handover interactions produced with a data-driven method trained on the 3HANDS dataset.
- We release the dataset to enable the community to create robust and reliable models of object handover with SRLs.

## 2 Related Work

### 2.1 Human-Human Handover

In recent years, the study of human-human handover [7, 58, 74] has gained attention due to its importance for improving human-robot interactions [19, 28] and collaborative systems [67, 70]. Past works have investigated a wide range of factors influential to handover activities. These include the use of interpersonal space [22], timing [20], handover context [4]. They further addressed factors related to the handover objects, such as their physical properties [4, 9, 12, 23], associated gripping dynamics [53], and transfer control of the object [74]. Other works have investigated giver and receivers' motions [26] to communicate intent before handover [73], as well as social bonding and shared goals [84]. Building upon these rich insights, significant advancement has been made in data-driven control methods for human-robot handover [3, 27, 32, 37, 85]. The data-driven approaches, which are trained on human-human handover data, have been shown to enable robots to better adapt to human behavior for smoother and more intuitive interactions [15, 68].

Several datasets have been developed to study human-human handovers, each varying in terms of setup, modalities, and object interactions. The HoH dataset [83] and the dataset by Khanna et al. [33] involve participants with a table between them, either seated or standing. HoH provides point clouds, while Khanna et al. include motion tracking along with handover forces. The H2O dataset [89], and the datasets by Kshirsagar et al. [36] and Chan et al. [10] involve participants standing at a comfortable distance, with variations in their sensor setups: H2O employs magnetic sensors and cameras to focus on hand dynamics, while the others

<sup>1</sup><https://hci.cs.uni-saarland.de/projects/3hands/>

utilize markered motion capture, RGB-D data, and multiple camera views. Carfi et al. [8] provide a more dynamic scenario, with participants freely moving toward each other in various handover contexts, incorporating multimodal data like motion capture, IMU, and videos. Lastly, Cini et al. [13] focus on grasps used during hand-object interactions.

To the best of our knowledge, existing datasets for human-to-human handovers are limited in that they focus on symmetric constellations where the giver and receiver face each other, and the handover occurs centrally in their shared interpersonal space. Furthermore, these datasets lack the implementation of an ongoing activity that is performed before and after the handover. To address these, we propose 3HANDS that is focused on an asymmetrical giver-receiver relationship in close peripersonal space while the primary user is also engaged in an activity. A comparison of the datasets is presented in the Table 1.

## 2.2 Human-Robot Handover Control

Human-robot handover tasks combine anticipation of human intent with path-planning algorithms to generate feasible and natural handover trajectories. Generated trajectories optimize safety, reachability, and timing, to ensure smooth and collision-free handovers. Models on natural reaching movement, such as the minimal jerk model [25] have been used for anticipating handover timing and location [39, 45]. Elliptic trajectory modeling [68] was proposed for making early and fast predictions on the movement of the collaborator and was shown to perform better than the minimal jerk model [11]. While classical modeling approaches provide fast computation times and hard constraints to ensure safety, they are prone to model inaccuracies and need more tuning effort from an experienced designer in custom scenarios.

In addition, classical modeling approaches are not well-suited to account for the subtleties and multimodality of naturalistic human interactions. Prior work on human-robot proxemics has highlighted the relevance of personal spatial zones for human-robot interaction [75, 81] and balanced physical distancing [6]. It has been shown that robots must follow societal norms of physical distancing to offer smooth and comfortable, rather than disruptive and threatening interactions [57]. Spatial invasion, due to inappropriate distances between the robot and the human, can result in discomfort and avoidance [42].

Data-driven or hybrid approaches are more capable of capturing subtle dynamics. The number of applications that rely on such methods is increasing as the availability of handover datasets improves. These approaches can be used to tune specific model parameters [54], make real-time predictions [52, 88], or control the entire process using generative models [61]. In this paper, we show that our 3HANDS dataset provides high-quality and detailed human pose data to enable the training of generative handover control architectures enabling naturalistic and fluid handovers without basic heuristic constraints. Furthermore, our work contributes to future analyses of personal space in human-robot interaction, an important area that is still in its infancy [41].

Handover control for supernumerary robotic limbs has received significantly less attention than for stationary robots.

Existing solutions either rely on human input by utilizing human redundant degrees of freedom [65] or use heuristic methods [16]. The lack of more data-driven approaches for wearable interfaces can be attributed to a lack of available datasets for training, imitating natural handover scenarios with an agent that resides in the user's personal space. We address this problem by contributing a comprehensive dataset focused on movement configurations that are specific to supernumerary robotic limbs.

## 2.3 Supernumerary Robotic Limbs

In recent years, wearable robotics have emerged as an expanding topic of study. These include supernumerary robotic limbs that augment users by providing additional extra limb-like robotic structures [86], prosthetics that replace missing body parts [30]; and exoskeletons that help to improve the physical performance of the user's existing limbs [87].

Supernumerary robotic limbs have been extensively researched, primarily in the robotics literature but also increasingly in HCI. Numerous structural configurations have been suggested by researchers for Supernumerary robotic limbs. For instance, prior work [79] proposed a forearm-mounted supernumerary robot, dexterous torso-mounted robotic arms [66], a shoulder-mounted extra arm for above-the-head work [48], or additional finger-like structures [44]. A pliable snake-shaped wearable robot featuring 25 degrees of freedom has been developed for highly adaptable application to the body in various geometric arrangements [1]. Various end-effectors are also suggested for SRLs [31]. Beyond physical assistance, SRLs are promising for virtual reality [5] and haptics, where wrist-worn [35] or waist-worn [2] robots can offer rich haptic feedback on multiple body locations. Another line of work investigates the important challenge of how to adapt an SRL to individual bodies of users and to individual body locations [90]. Key directions include creating customized SRLs by assembling modular hardware building blocks [43] or using motion capture, digital design, and optimization algorithms to digitally customize a device design for computational manufacturing [64]. A central question involves how to control the motion of SRLs. To manage the motion trajectories of the SRLs, researchers have investigated the interactions between the user and the device [56]. This is a hard challenge because, when operating an SRL, the user's body is frequently occupied with a primary manual activity, restricting conventional touch or gesture-based interaction. One line of inquiry centers on robot planning, which employs activity recognition to autonomously steer the robot toward a goal that negates the need for direct human interaction [49]. Another line of inquiry uses remapping of body motion, where degrees of freedom in body movement that are not required for a specific task are remapped to control the SRL. For instance, mapping a user's foot movements to robotic arms can be a promising technique for intuitive and flexible real-time control [65]. Other approaches proposed using the back of the hand [40], the pinky finger [46], or capturing muscle movements with EMG [51] for controlling an SRL. Our work contributes to interactions with SRLs by proposing to

Dataset	Cini [13]	Chan [10]	Carfi [8]	Kshirsagar [36]	H2O [89]	HOH [83]	3HANDS (ours)
Human-human spatial zone	-	Social	Social/public	Personal/social	-	Personal/social	Close intimate/intimate
Activities	✗	✗	✗	✗	✗	✗	12
Interactions	1734	1200	288	240	1200	2720	946
Unique participant pairings	17	10	18	24	40	40	12
Markerless	✗	✗	✗	✗	✗	✓	✓
Cameras	1	8	1	2	5	8	41
Full body 3D skeleton	✗	✗	9 joints	13 joints	✗	✗	69 joints
Hands 3D skeleton	✗	✗	✗	✗	✗	✗	21 joints per hand
Objects	17	20	7	5	30	136	3
Experimental Validation	✗	✗	✗	✗	✓	✓	✓
Setting	-	Standing, Freely moving	Standing, Freely moving	Standing, Face-to-face	Standing, Face-to-face	Seated, Face-to-face	Standing-seated, Asymmetric
Suitable for SRLs	✗	✗	✗	✗	✗	✗	✓

**Table 1: Comparison of 3HANDS with prior human-human handover datasets. Human-human spatial zones are inferred based on Lambert’s definition of spatial zones [38].**

learn subtle and nuanced motion dynamics from pairs of interacting humans.

### 3 Dataset

In this section, we detail on the 3HANDS dataset, an extensive collection of motion patterns originating from two persons engaging in an object handover. It comprises detailed motion data of more than 946 interactions where the primary person is performing 12 daily activities while the second person is enacting a hip-mounted third arm that hands over and takes back objects to assist the primary person during the daily activity. The decision to use a hip-mounted SRL in 3HANDS is based on its popularity in related work and its ability to minimize interference with the user’s natural active space compared to other common SRL mounting locations. Additionally, it enhances safety by keeping the SRL away from sensitive areas such as the head and face. By recording interactions that were intuitively performed by two interacting humans, the dataset captures the specific and mostly implicit requirements of operating in intimate personal space as well as the interpersonal dynamics of object handover during a primary activity.

#### 3.1 Apparatus and Captured Motion Data

We recorded the participants using the markerless optical motion capture system *Captury*<sup>2</sup>, which is based on the skeleton tracking approach of Stoll et al. [72] with additional hand tracking and a comparable average range of error of 8.79 mm compared to the marker-based Vicon system (cf. [24]). The allocentric setup uses 41 time-synchronized RGB cameras mounted at the walls and ceiling, each recording at a resolution of 2056 × 1504 pixels with 25 Hz framerate. This multiview motion capture effectively minimizes data loss caused by occlusions, as occluded joints are likely visible in other views. As extreme occlusions could cause challenges similar to marker-based systems, our manual verification confirmed the motion quality without any instances of mistracked joints. They capture the motion of multiple persons simultaneously, in an area of 7 × 6 m. In a *markerless* motion-capture setup, the participants are not required to wear body suits or stick optical markers on their

bodies. This *non-invasive* capture setup allows the participants to freely and naturally perform the desired tasks with no restrictions on the kind of motions they can exhibit. Additionally, the system also provides tracking of finger joints, thereby allowing us to capture fine-grained handover. Not having markers facilitates better capture of such finger articulations as it is typically difficult to attach and label markers on the fingers.

The setup provides rigged skeleton data of both interacting participants, including their hand poses (21 joints for each hand). As human joints have limits on the articulation angle and not all joints rotate along all three axes, the mocap system defines the skeleton as a kinematic tree of 107 *Degrees-of-Freedom* (DoF). Each DoF represents the axis-aligned rotation of a joint along a specific axis defined in the local coordinate system. The DoFs are also assigned individual limits on the maximum and the minimum articulation based on statistical data. These skeleton DoFs can be transformed into body joint rotations represented using Euler angles (or quaternion). Further, we perform a Forward Kinematics operation on the joint angles to recover the 3D positions of each body joint. In total, our skeleton definition comprises 69 joints and 107 Degrees of Freedom (DoF).

We also recorded the audio of the spoken instructions provided by the primary participant using an omnidirectional neckband microphone attached to the participant. In order to synchronize the audio with the captured motion, we ask the participant to clap three times at the start and the end of the recording sequence. The peaks at the audio channel and clap moments of the hand joints are then aligned to achieve synchronization.

#### 3.2 Activities

Since we aim to capture a broad spectrum of motion patterns, we let participants perform 12 manual activities, representative of everyday activities that frequently require object handovers. In order to select activities that represent a broad set of tasks, we systematically select such activities that provide a large coverage along the following parameters of motion patterns: 1) **Height** at which the action occurs relative to the user’s body; we include tasks at the level of the user’s *torso* and *head*. 2) **Distance from the body**: we vary activities carried out *on-body* (these require

<sup>2</sup><https://captury.com>

particularly careful motion to avoid uncomfortable or even hurtful encounters) and *mid-air* activities carried out at a certain distance in front of the body. 3) **Motion range**: we distinguish between *small* motion range, where hands stay largely at the same location for performing dexterous tasks (e.g., adjusting a small picture at the wall), *medium* range of motion (e.g., hammering a nail), and *large* range of motion (e.g., painting a large wall).

Based on the parameters defined above, we select 12 everyday manual activities that each cover a unique combination in this  $2 \times 2 \times 3$  parameter space. Table 2 depicts the set of activities. We provided the participants with authentic props for each task to enhance the realism of performing the activities. For instance, a hammer was provided for hammering, a washcloth for washing the torso, etc. Additionally, we recorded a neutral pose where users were instructed to comfortably rest their arms while standing still.

### 3.3 Task and Procedure

**3.3.1 Roles and Spatial Setup.** One participant takes the role of the user (called the *primary participant*), and the other one acts as the serving robotic arm (called the *robot participant*). Pairs were instructed that the robot participant should aim at assisting the primary participant to the best level possible in handing over and taking objects, while the primary participant should focus on the primary activity and not care about the robot participant. Contrary to previously introduced handover datasets [8, 10, 13, 36, 83, 89], our participant pairs do not face one another. Instead, we intentionally arranged the setup to resemble a hip-mounted SRL on the dominant hand's side. Therefore, the robot participant was asked to sit on the dominant hand side of the primary participant where the shoulder of the robot participant is at the hip level of the primary participant, facing the primary participant's hips at a slight distance (approx. 20cm), so as to not block the primary participant's elbow while performing the activity. The primary participants were standing and wore glasses that shielded their peripheral view on the lower right. This shielded the robot participant's face from their peripheral view, enabling them to focus on their activity and avoid communicating through eye contact. We illustrate the arrangement of the participants in Figure 2.

The height of the stool for the robot participant is adjusted such that the robot participant's shoulder is aligned with the height of the primary participant's hip level. For the activities that required to be performed on a wall, we provided a wall-sized fixed acrylic panel in order to not block the camera views.

**3.3.2 Handover Task.** The experimenter first communicated the general instructions by playing a voice recording. After a short trial run, the pairs then performed the following handover task for each of the 13 activities (12 + 1 neutral pose, activities were performed in randomized order): After hearing a beep sound, the primary participant is performing the activity with the provided prop object, standing upright. The robot participant's right arm is in a resting position (hanging down), holding the handover object. Next, the primary participant initiates a handover at any preferred time. The primary participant is free in the modality and way they

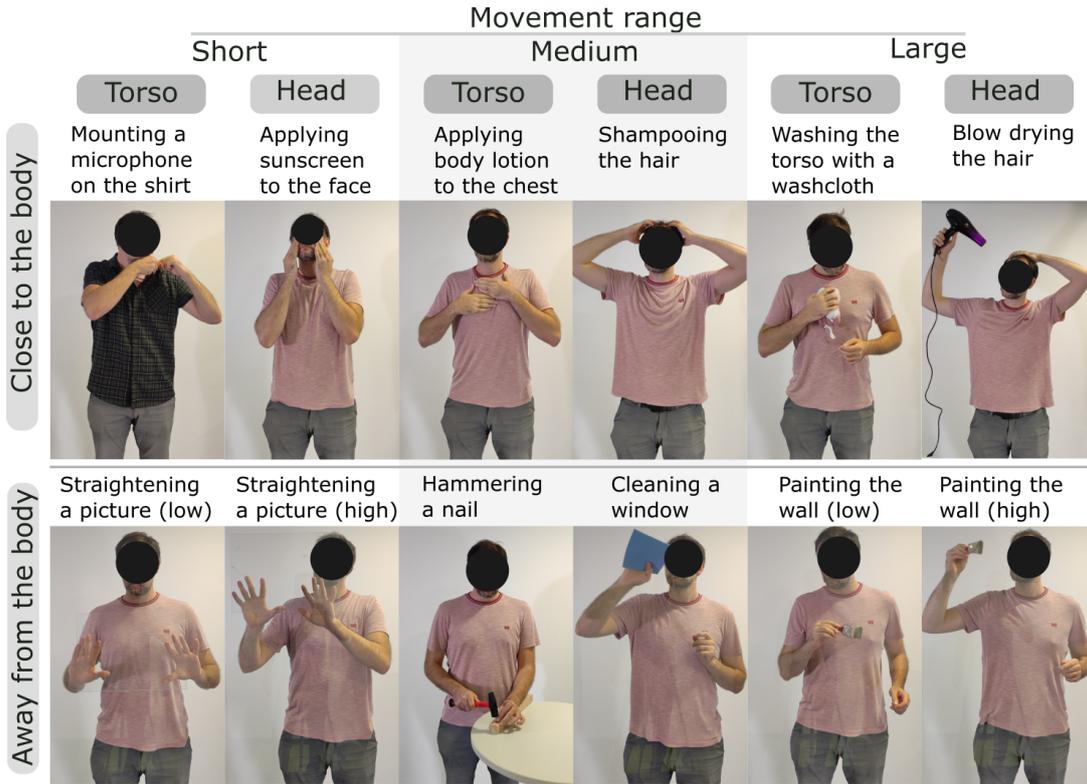


**Figure 2: Setup of the asymmetric handover task. Left: the primary participant was standing and performing the primary activity, while the second participant enacted a robotic arm for handing over an object. Right: we generate rigged skeletons of both humans, including their articulated hands.**

would want to make the robot participant aware of their intention for handover. Then, the robot participant starts handing the object over to the primary participant and then goes back to the resting position. The participants were instructed to perform this motion in a way they considered natural. Once the object is handed over, the primary participant briefly mimics using the object. At any preferred time, the primary participant then signals the robot participant to take away the object. The robot participant's right arm starts moving, takes the object, and returns to its resting position.

Since we solely focus on the motion patterns and not on the specifics of the grasp, we assigned a fixed handover object per activity that is chosen among three spherical objects with diameters 2.5, 4, and 6cm, assigned with relevance to handover objects in the scenarios.

**3.3.3 Trials.** The participants repeated the handover three times for each activity. After this process, the pair reversed roles and fully repeated it again, resulting in 156 trials (12+1 activities x 2 hand-to/take-away x 3 repetitions x 2 reversed roles). In 16 instances, participants have performed 4 instead of 3 repetitions. Data from 7 trials had to be discarded due to the primary participant looking at the robot participant's face or the robot participant's right hand not waiting in the rest pose. In summary, the dataset contains 946 captured interactions. For one pair of participants (including reversed roles), the whole capturing session took approximately 90 minutes.



**Table 2: The set of activities captured in the dataset. Transparent acrylic sheets were used for the wall and the picture frame to avoid visual occlusions.**

### 3.4 Participants

We recruited 12 participants (6 male, 6 female, 0 non-binary), aged between 21 and 33 years old. All participants reported being right-handed. They received a monetary compensation. Participants conducted the data capture in pairs and reversed roles to double the number of unique pairings. Since our setup requires operating in the intimate peripersonal space around the body, we opted for recruiting only couples who are in a stable relationship. To ensure that the arm’s length while sitting is sufficient to reach the location for object handover, we only included couples whose difference in height was not more than 20 cm.

### 3.5 Data Processing

The output of the preprocessed data is the motion data of the interacting participants. It includes the motion files (in BVH and FBX format) together with time-synchronized raw video of all 41 cameras and the audio recording.

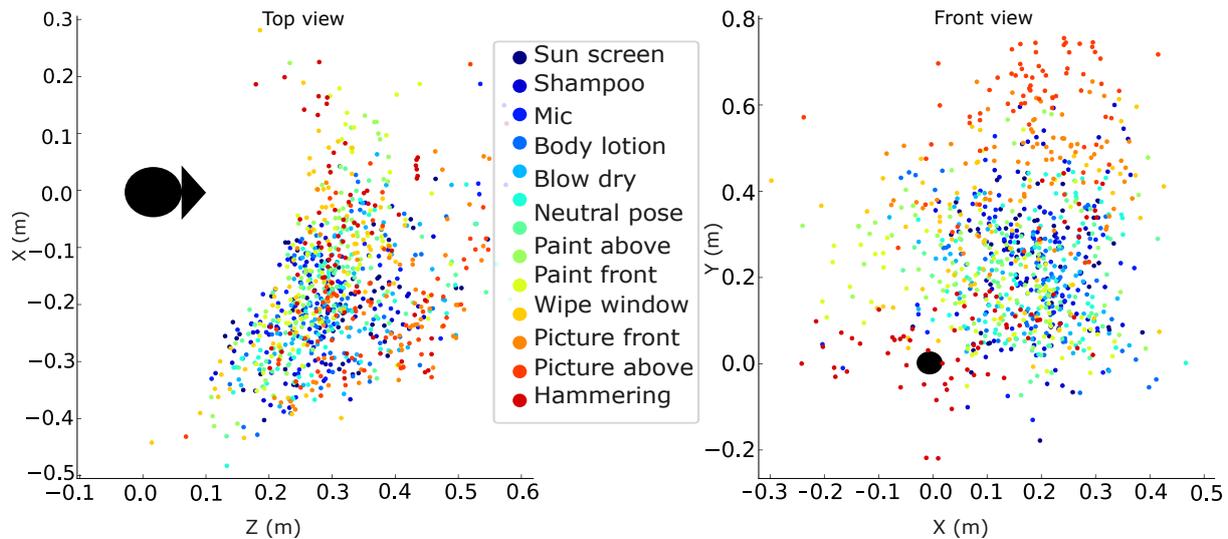
We then manually annotated the Ground Truth (GT) by marking the frames that belong to a handover activity. We define a handover activity to begin when the robot user starts moving; it ends when the robot user’s hand has returned to its resting position after the object has been handed over. Each handover frame is annotated with the correct label of the specific handover task (give to or take away). In every frame of a handover, we furthermore labeled whether the object was in the primary participant’s hand

or in the robot participant’s hand. The timestamp for every valid frame in its relevant handover segment is also added to the data (time data). This can be used later to provide temporal information about the current status of the interaction with the model.

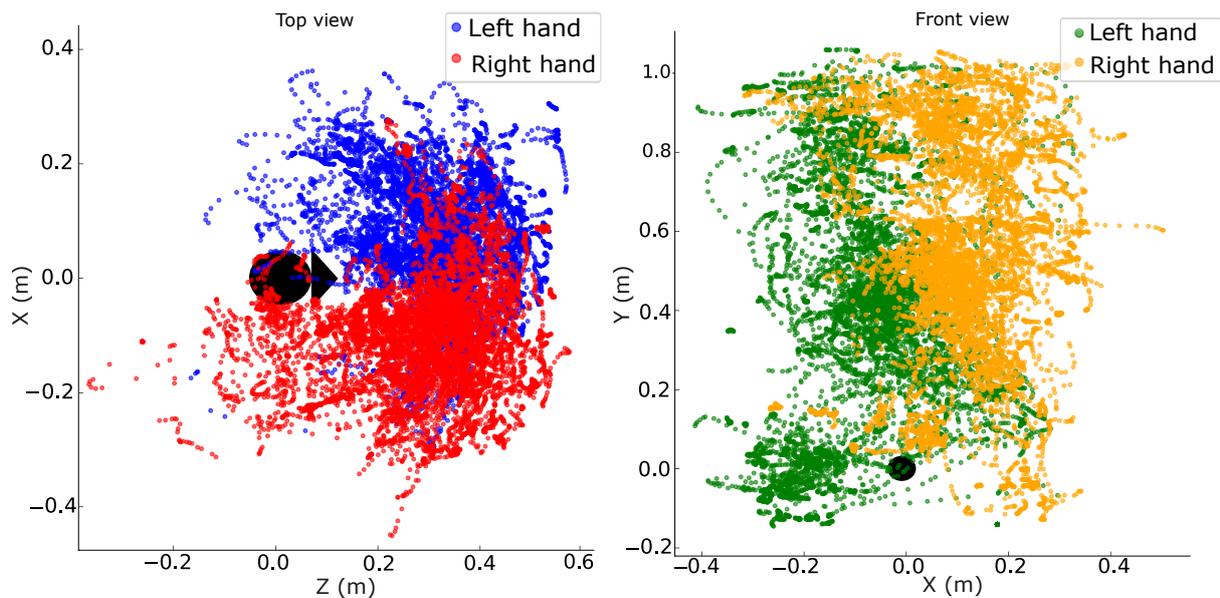
For each frame of a handover, after marking the start and end times of the GT segments, we store the 3D rotations (in parent-relative coordinates) and calculate and restore the 3D positions (in global coordinates) of each joint of both participants for all the valid frames within the sequence. We also include the stored rotation and position joint values in the dataset (CSV format). The 3D joints’ positions’ calculations are based on the root joint’s location and rotation (hip joint of each participant), following the skeletal hierarchy and using the local rotations and segment lengths specified in the motion files. We also provide the transcription of the associated audio files including verbal commands and reactions for handovers, as well as the annotation of the activity and handover tasks. It is worth noting that although we do not use the robot participants’ full-body motion information in our experiments (Section 4), yet we release them as additional annotations for the community to work on.

### 3.6 Data Analysis

We conducted an initial analysis of the dataset to identify significant patterns in the motion data. The average duration of a handover across all activities was  $2.244 \pm 0.854$  seconds.



**Figure 3: Distribution of the locations where the object was handed over between participants. The points are presented in the user's hips coordinate system. (left) shows the distribution from the top view (head at origin, facing towards right) and (right) from the front view (hip at origin, user facing inwards the plane).**



**Figure 4: Distribution of the palm over the entire dataset for performing the 12+1 activities. The points are presented in the user's hips coordinate system. (left) shows the top-view (head at origin, facing towards right), (right) the frontal view (hip at origin, user facing inwards the plane). The color encodes the left and right hands. The unit is meters.**

Another notable aspect of our dataset is where exactly the object was handed over between participants depending on the activity. Figure 3 visualizes the distribution of handover locations in the user's hip coordinate system, color-coded for each activity. The distribution of handovers from the top view (see Figure 3 left) shows that the distribution mainly extends to a hemispherical region of up to approx. 0.5 m to the primary user's front and approx. 0.45 m to their right

side. Interestingly, it exhibits a distinct skew towards the primary participant's right side, influenced by the positioning of the robot participant on this side. The front view (Figure 3 right) shows that handovers were primarily performed in an area ranging from hip-level to approx. 0.5 m above hip level, while some handovers, primarily for activities performed at the head level, extend up to approx. 0.8 m above hip level.

Moreover, we were interested in the distribution of motion across all activities because we hypothesized that the varied conditions under which the handovers were performed would result in a wide range of motion patterns, reflecting the flexibility and adaptability of human motor behavior in response to different task demands. As depicted in Figure 4, the distribution of the primary user’s palm position with regards to the user’s hip joint throughout the activities reveals extensive coverage across the entire space in front of the user. The top view (Figure 4 left) and frontal view (Figure 4 right) both show that the primary user utilized a broad range of motion from left to right and head to hip, encompassing nearly all reachable areas, which supports our hypothesis of a large distribution of motion.

## 4 Validating the Dataset with Models

This section showcases our dataset’s usability for generating handover trajectories and predicting key handover characteristics through the following task settings:

- (1) First, we show that our dataset enables training of generative models which **synthesize handover trajectories** of an SRL in a human-like manner. To this end, we train a conditional variational auto-encoder on the complete handover trajectories of the robotic participant, conditioned on the full-body motion of the primary participant. In effect, this task learns *how* the robotic arm should move.
- (2) Secondly, we show that with our data, we can train a model that predicts the **locations of handover** aligned with the actual handover locations. We train a conditional variational auto-encoder to predict the potential handover position and orientation at any given time on the trajectory. This task informs *where* the robotic arm should move to.
- (3) Lastly, we validate that our dataset contains the vital information for training a binary classifier that predicts **when a handover occurs** by only observing the primary user’s motions. This informs *when* the robot should start to move for a handover.

For all the tasks described above, we provide details on the data processing steps, the model, and the training process and report technical evaluation results, where the model predictions are compared against the testing data gathered from human participants. We also identify the most influential joints to make training and system deployment more efficient.

### 4.1 Generating the Trajectory of a Handover

A key aspect that we aim to demonstrate in our dataset is allowing models to generate human-like handover trajectories, which is a challenging task due to the highly variable and high dimensional nature of the human body’s motion space. Specifically, our goal is to generate motion trajectories from the starting point to the handover position, dynamically accounting for the posture changes of a user during the handover.

#### 4.1.1 Data Processing.

*Motion representation.* We define human motion as time series data of sequential human body poses with timeframe  $T$ . At any given timestamp  $t$ , our dataset contains the positional and rotational data for all joints of both the primary human participant and the robot participant. The data processing follows standard methods commonly employed in motion generative and predictive models (e.g., [47]). The joint’s position ( $j_p$ ) is represented in the rigged character’s root coordinates. Each joint’s rotation ( $j_r$ ) is represented in its local Euler angles. We normalize the poses by translating and rotating such that the root joint (hip) is positioned at the origin of the world coordinate system and skeletons are oriented uniformly in the same direction. Finally, to maintain a continuous representation of joint rotations, we project the 3D rotational data into a continuous 6D space ( $j_r \in \mathbb{R}^6$ ), a widely used technique [91].

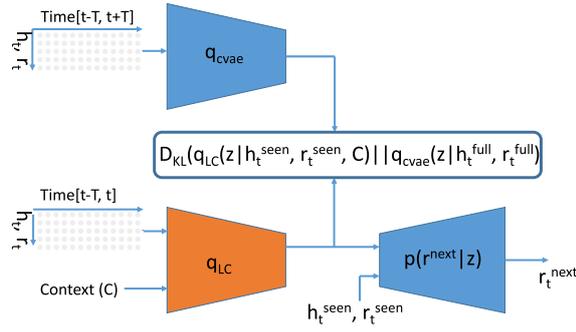
The pose of the primary participant at timestamp  $t$  is expressed as a tuple of joint positions and local rotations,  $h^t = [j_p^t, j_r^t]$ . The human motion between timestamp  $t_1$  and  $t_2$  then is defined as the sequence of poses between  $t_1$  and  $t_2$ :  $h^{[t_1:t_2]} = [h^{t_1} : h^{t_2}]$

Similarly, we define the robot user’s pose as  $r^t = [EE_p^t, EE_r^t]$  where  $EE_p$  and  $EE_r$  are the end-effector (right hand of the robot participant) position and rotation respectively. The robot motion is then  $r^{[t_1:t_2]} = [r^{t_1} : r^{t_2}]$ .

*Input and output data.* There are three inputs to our model: the primary participant’s full-body motion, the robot participant’s end-effector motion, and the handover state ("handing over", "taking back," or "idle"). The motions of the primary and robot participants are symmetric time windows looking back  $T$  timestamps into the past and  $T$  timestamps into the future. We can describe such input data of the primary participant as  $h^{[t-T:t+T]}$  and the robot participant as  $r^{[t-T:t+T]}$ . We set  $T = 25$ , corresponding to a duration of one second. The output of our model is the generated 3DoF next position of the robot end-effector mapped to the robot participant’s right hand from the 3HANDS.

*Train and test data separation.* Our data comes from human subjects, where each participant’s data is likely to contain unique patterns or behaviors. We perform a train-test split on the participant level to ensure the test data is completely unused for training. Specifically, we randomly select two participant pairs (i.e., 4 participants’ data) as the test set, which are not involved in model training. The remaining 8 participants’ data is utilized for training, and not involved in testing at all. This data separation is consistent throughout all the following models.

*4.1.2 Model.* Conditional Variational Autoencoders (CVAE) [69] are one of the most popular architectures widely used for generating motions or poses [14, 18, 47, 63, 71], with an encoder-decoder structure. The encoder compresses high-dimensional pose data into a smooth, continuous latent space ( $z$ ), while the decoder generates the next pose based on this representation and conditional inputs. By modeling the probabilistic distribution of the data, CVAEs can generate diverse and realistic motion sequences through sampling.



**Figure 5: Overview of our proposed model architecture for generating the trajectory of a handover based on the motion dynamics encoded into the model's latent space.**

Among CVAE-based models, our approach builds upon the Motion Variational Autoencoder (MVAE) [47]. We chose MVAE as the basis for our model because of its demonstrated capabilities in autoregressive motion generation making its output robust against the highly variable motion inputs. MVAE utilizes a Mixture of Experts (MoE) architecture in the decoder, refining predictions by incorporating multiple specialized networks for different motion aspects.

Building on MVAE, we propose **SVAE (CVAE for SRL)**, a model specifically designed for SRL's motions in handover processes. The model architecture is illustrated in Figure 5. While MVAE and SVAE share the same foundational framework, SVAE addresses two novel challenges: generating SRL motion by considering both the user's and the robot's movements and adapting to varying handover states, which MVAE does not account for. A key enhancement in SVAE is the integration of attention mechanisms across its components enabling the model to better capture the relevant temporal and spatial aspects of the input motion data. The *encoder* in SVAE, similar to MVAE, compresses high-dimensional data into a latent space, but it processes motion data over a time window that includes current, past, and future timesteps, unlike MVAE, which only handles current and past data. The *decoder* generates the next pose by sampling from the latent space, but in contrast to MVAE, our model is conditioned on both the robot's and human's observed motions. SVAE keeps the *Mixture of Experts (MoE)* architecture for the decoder, utilizing six expert networks and a gating mechanism. Additionally, our model introduces a *latent controller (LC)*, which aligns the latent space with the specific handover state ("handing in," "taking away," or "idle"). This context-aware layer provides enhanced control, allowing SVAE to manage the timing and variability of human-robot interactions during handovers, an essential feature not present in MVAE. The latent controller also incorporates an attention mechanism to adaptively align the input motions and handover state with the latent representation learned by the encoder.

Here, we provide the loss function based on the Evidence Lower Bound (ELBO) to train the SVAE :

$$ELBO_t^r = \mathbb{E}_{q_{LC}} [\log p_{\theta}(r^{next} | z, h^{seen}, r^{seen})] - \beta \text{KL} \left( q_{LC}(z | h_t^{seen}, r_t^{seen}, C) || p(z | h_t^{full}, r_t^{full}) \right)$$

where  $t$  represents the current timestamp.  $h$  and  $r$  refer to the primary participant and robot participant, respectively. The loss function has two main components. The first term is the reconstruction loss  $\mathbb{E}_{q_{LC}} [\log p_{\theta}(r^{next} | z, h^{seen}, r^{seen})]$ . It measures how well the decoder can predict the next SRL pose  $r^{next}$ .  $z$  is the latent variable encoding the motion information,  $h^{seen}$  represents the observed human motion, and  $r^{seen}$  represents the observed SRL motion up to the current timestamp. This term ensures that the model generates accurate and contextually appropriate motions. The second term,  $KL$ , is the Kullback-Leibler (KL) divergence, which regularizes the latent space by aligning the learned posterior distribution  $q_{LC}(z | h_t^{seen}, r_t^{seen}, C)$  with a prior distribution  $p(z | h_t^{full}, r_t^{full})$ . Here,  $h_t^{full}$  and  $r_t^{full}$  represent the full human and SRL motion data across the time window, and  $C$  denotes the handover state ("handing in," "taking away," or "idle"). The parameter  $\beta$  controls the trade-off between the two components in the ELBO. Balancing the model's ability to reconstruct accurate motions and maintaining a well-structured latent space that generalizes effectively, we used  $\beta = 0.1$  in our setting.

**4.1.3 Training.** Our training process is divided into two main stages. In the first stage, we train the SVAE model for 140 epochs. Following that, we use the trained SVAE to train the complete pipeline, which includes both the SVAE and the LC encoder, for 250 epochs. We begin with training the SVAE model for 10 epochs using only the reconstruction loss, after which we introduce KL divergence loss into the loss function.

$$\mathcal{L}_{SVAE} = \mathcal{L}_{rec} + \beta \cdot \mathcal{L}_{KL}(z_{SVAE}, \mathcal{N}(0, I))$$

We employ the adaptive moment estimation (ADAM) optimizer, with a learning rate that decays from  $10^{-4}$  to  $10^{-7}$ , starting to decay at the 50<sup>th</sup> epoch and continuing throughout the training.

In the second stage, we align the output of the LC encoder with the learned latent space of the SVAE while freezing the weights of the encoder in SVAE. For the first 50 epochs, we minimize a loss based on the KL divergence between the learned latent space  $z_{SVAE}$  and the latent of the LC encoder  $z_{LC}$ . After that, the reconstruction error is incorporated into the loss, and training continues for an additional 50 epochs. The loss function for the training process is:

$$\mathcal{L}_{LC, SVAE} = \mathcal{L}_{rec} + \beta \cdot \mathcal{L}_{KL}(z_{SVAE}, z_{LC})$$

We trained our model using reconstruction loss based on the  $l_2$  distance of the 3DoF of the robot's end-effector ( $y$ ), which corresponds to the right hand of the SRL participant from the dataset ( $y$ ).

We apply scheduled sampling during training, where the model's output is fed back as input for autoregressive generation over  $l = 10$  consecutive steps. The probability of using autoregression,  $p$ , increases from 0 to 1 over 50 epochs, after which the model is fully autoregressive ( $p = 1$ ). The

ADAM optimizer is used again in this stage, with learning rate settings similar to the first stage.

**4.1.4 Testing Results.** We generate trajectories in two forms. One is non-autoregressive: we take the seen motion sequence of the test dataset to generate a single next position ( $\hat{y}_t$ ) and evaluate it against the ground truth position in the test data ( $y_t$ ). The other is autoregressive: we provide the model with an initial starting position, which then continuously generates the next positions based on the very previously generated 25 robot positions.

To quantify the quality of the generated trajectories, we report the Mean Absolute Error (MAE) of pairwise comparison between the generated and the ground-truth trajectories:  $MAE = \frac{1}{\eta} \sum_{j=1}^{\eta} \frac{1}{N} \sum_{i=1}^N |y_t - \hat{y}_t|$ , where  $N$  is the length of the trajectories normalized by the number of trajectories  $\eta$ .

Table 3 shows the MAE errors. Our model generates handover trajectories with MAEs ranging between 2.10–2.71 cm in the non-autoregressive setting. With autoregression, MAEs range between 10.42–23.85 cm. We observe that our model shows strong performance in the pairwise comparisons with the ground-truth in the non-autoregressive setting. The error increases in the autoregressive setting, which is expected as errors accumulate. It is important to highlight a key distinction between our motion generation task and previous related work (e.g., [47]). In most prior approaches, motion generation is based on observations of the same actor. In contrast, our model generates the motions of one actor (the robot user) based on the observations of a different actor (the primary user). This fundamental difference introduces additional complexity. Both actors interact in a continuous real-time feedback loop, where the motion of one most likely directly influences the motion of the other. This dynamic interplay cannot be fully accounted for in the autoregressive setting. It is to be assumed that in interactive real-world deployments, the error will be lower, as the primary user would adapt their motion trajectory to the robot’s trajectory. As this is a novel research question, future works should investigate more advanced models to further mitigate the error. One potential idea is to employ reasoning models, which infer the primary users’ intents and use that to condition or correct the robot’s motions.

For a qualitative impression of the generated trajectories, we refer the reader to the Video Figure where we show several representative trajectories generated for different activities. Results show that our model generates plausible trajectories that are in keeping with the key characteristics of naturalistic human handover motion in close peripersonal space.

To quantify the individual joints’ importance for generating motions, we employ a gradient-based sensitivity analysis [34, 55]. A higher gradient indicates this input feature has a higher influence on the output. The result is shown in Table 4. Our findings suggest that, on average, the positions of the "left shoulder", "right elbow" and "left hand" joints are most influential, while the rotations of the "Face", "right elbow" and "right hand" offer the most decisive rotation information for generating the handover trajectory. The neck’s position and the back’s rotation are the least decisive features.

Activity	MAE W/O autoreg. (cm)	MAE W/ autoreg. (cm)
Mount a mic	2.55 ± 2.61	15.22 ± 16.52
Apply sunscreen to face	2.70 ± 3.10	14.13 ± 15.95
Apply body lotion to chest	2.60 ± 2.34	15.09 ± 16.67
Shampoo hair	2.57 ± 2.25	12.77 ± 15.73
Wash torso with washcloth	02.55 ± 2.33	15.03 ± 14.51
Blow dry hair	2.55 ± 2.32	15.77 ± 15.58
Straighten a pic (low)	2.34 ± 2.03	22.28 ± 24.55
Straighten a picture (high)	2.60 ± 2.84	23.85 ± 28.23
Hammer a nail	2.10 ± 1.80	10.42 ± 8.71
Clean a window	2.71 ± 2.50	14.79 ± 15.85
Paint the wall (low)	2.51 ± 2.43	10.72 ± 10.54
Paint the wall (high)	2.59 ± 2.15	14.15 ± 16.59
Neutral pose	2.60 ± 2.50	12.48 ± 13.88

**Table 3: Results for generating the trajectory of a handover. Mean absolute error in meters (m) is reported.**

JIR	Joint Position	Joint Rotation
1	Left shoulder	Face
2	Right elbow	Right elbow
3	Left hand	Right hand
4	Chest notch	Left hand
5	Head	Right wrist
6	Right shoulder	Upper back
7	Right hand	Left clavicle
8	Back	Head
9	Right clavicle	Right clavicle
10	Face	Left elbow
11	Upper back	Left wrist
12	Write wrist	Right shoulder
13	Left clavicle	Neck
14	Left elbow	Chest notch
15	Left wrist	Left shoulder
16	Neck	Back

**Table 4: Joint Importance Rank (JIR) for handover trajectory generation, indicating the relevance of an individual joint’s position and rotation generating motions.**

## 4.2 Generating the Region of Transfer

The 3D location where the object is a crucial characteristic in handover motions, and so is the rotation of hands at the time of handover. Such information regarding the position and rotation of the handover is also known as the “region of Transfer” or ROT [83]. Given that this ROT is the final product of a full trajectory, ideally, it can be predicted by the observed segments of the trajectory. In this subsection, we show that our dataset contains detailed motions that enable ROT prediction.

**4.2.1 Data Processing.** Following Wiederhold et al. [83], we define the handover coordinate (location) as the midpoint and the orientation as the direction of the axis that passes through the palms of the primary participant and robot

participant at the time of handover. At each timestamp  $t$  during the handover, the input to our model is the primary user’s motion  $h^{[t-T:t]}$  and the robot user’s motion  $r^{[t-T,t]}$  over the past time period  $T=25$  (1 sec).

We acquire the ground truth handover coordinate and rotation as the 3D mid-point of user-SRL hands and the 3D vector pointing from the giver’s hand to the receiver’s hand that shows the orientation of the user-SRL hands at the transfer time stamp. The output of our model is the 6 DoF-generated ROT for the current motion of the user-SRL.

**4.2.2 Model.** The generation of the RoT is very aligned with generating the trajectories in the previous subsection, as both are taking the primary user’s motions as input, however, the main difference is the timing of the output. While the trajectory generation relies on the autoregressive generation of the next positions sequentially, the RoT generation does not have this constraint. Therefore, we employ a conditional variational autoencoder (CVAE) [69] for this task because its capability in pose generation has been demonstrated in previous work [18, 63]. The encoder  $\phi$  in our CVAE encodes the input motions into the latent value  $z$ . The decoder  $\theta$  in our model samples from the latent distribution  $z$  and, conditioned on the human and the SRL motions, generates the 6 DoF ROT.

**4.2.3 Training.** We train the pipeline for 250 epochs with an adaptive moment estimation (ADAM) optimizer, with a decaying learning rate from  $10^{-4}$  to  $10^{-7}$ . The model is trained with  $l_2$  distance for position and orientation of the RoT.

$$\mathcal{L} = \frac{1}{2} \left( \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 + \|\hat{\mathbf{q}} - \mathbf{q}\|_2^2 \right)$$

where  $\hat{\mathbf{p}}$  and  $\mathbf{p}$  are the generated and ground truth 3DoF positions respectively, and  $\hat{\mathbf{q}}$  and  $\mathbf{q}$  are the generated and ground truth 3 DoF orientations.

**4.2.4 Testing Results.** We examine the performance of the model in predicting the 6 DoF features of the RoT at each time stamp during the handover process by observing the motions from the past 1 second. Table 5 reports the mean absolute error (MAE) for the 3 DoF position and the mean Euler angle error (MEAE) for the orientation of RoT. The results show that the model achieves MAEs that range between 4.02cm and 8.04cm, while the achieved MEAE is between 0.0002 and 0.004 radians. These relatively low errors indicate that our dataset captures sufficient data to allow for predicting the Region of Transfer.

Furthermore, we investigate the importance levels of individual joints of the primary user’s and the SRL’s motions for predicting the RoT information. The results are shown in Table 6. The left elbow’s position and rotation are reported to be the most influential feature impacting the whereabouts of the RoT.

### 4.3 Predicting the Timeframe of Handover

We demonstrate our dataset’s capability to predict the moment the primary user wants to initiate a handover, just from observing the primary user’s motions. Similar to how humans use implicit cues without verbal expressions in

Activity	MAE (cm)	MEAE (rad)
Mount a mic	6.24 ± 2.82	0.0207 ± 0.0142
Apply sunscreen to face	7.37 ± 3.67	0.0268 ± 0.0196
Apply body lotion to chest	6.01 ± 2.85	0.0211 ± 0.0186
Shampoo hair	8.04 ± 3.14	0.0165 ± 0.0121
Wash torso with washcloth	6.71 ± 3.12	0.0221 ± 0.0125
Blow dry hair	7.45 ± 3.01	0.0240 ± 0.0176
Straighten a pic (low)	4.88 ± 2.45	0.0076 ± 0.0054
Straighten a picture (high)	5.69 ± 3.41	0.0092 ± 0.0088
Hammer a nail	4.02 ± 2.49	0.0099 ± 0.0067
Clean a window	6.13 ± 3.26	0.0159 ± 0.0107
Paint the wall (low)	4.20 ± 2.13	0.0132 ± 0.0097
Paint the wall (high)	6.41 ± 3.83	0.0179 ± 0.0124
Neutral pose	4.72 ± 2.57	0.0102 ± 0.0100

**Table 5: Test results for generating the region of transfer: mean absolute error of the generated positions (left) and rotation angles (right).**

JIR	Joint Position	Joint Rotation
1	Left elbow	Left elbow
2	Chest notch	Chest notch
3	Face	Left shoulder
4	Head	Right shoulder
5	Right wrist	Right elbow
6	Right clavicle	Right wrist
7	Left wrist	Neck
8	Left clavicle	Upper back
9	Left hand	Face
10	Right elbow	Left wrist
11	Left shoulder	Left hand
12	Right hand	Right clavicle
13	Back	Head
14	Upper back	Left clavicle
15	Right shoulder	Back
16	Neck	Right hand

**Table 6: Joint importance rank (JIR) for generation of Region of Transfer, indicating the relevance of an individual joint’s position and rotation.**

human-to-human handover [82], we demonstrate that our dataset encapsulates such implicit cues and thus allows for training a model for prediction. We define this problem as a binary classification problem, where the model is trained to predict whether a handover is currently ongoing or not.

**4.3.1 Data Processing.** As "handover", we define the sequence of frames that begins when the robot user starts moving, and that ends when the robot user’s hand has returned to its resting position after the object has been handed over. Our dataset comprises ground-truth annotation with a binary variable  $y$  that indicates for each frame whether it belongs to a "handover" or not.

At each timestamp,  $t$ , the input to the model consists of a sliding window of the user’s motion data over a time window of length  $T = 25$  (1 sec) previous to the current

timestamp  $h^{[t-T:t]}$ . The model’s output at each timestamp  $t$  is a continuous float value representing the likelihood of a handover being in progress at  $t$ . We use thresholding to convert this likelihood into a binary classification result: any value greater than 0.6 is considered a "handover".

**4.3.2 Model.** We employ a model composed of a 3-layered fully connected neural network with 128 nodes in each layer. For the activation functions, the first two layers are followed by the ELU function, and the last layer is followed by the Sigmoid function after the last linear layer, to ensure the output is bounded to  $[0, 1]$ .

**4.3.3 Training.** We trained our model on all instances of handover in the train data, regardless of how the participants had communicated their handover intent. We used the same optimizer type (ADAM) and decaying learning rate ( $10^{-4}$  to  $10^{-7}$  as in our previous experiments. We trained the model for 500 epochs with the binary cross-entropy loss function.

$$\mathcal{L} = -\frac{1}{N} \sum_{t=1}^N [y^t \log(\hat{y}^t) + (1 - y^t) \log(1 - \hat{y}^t)]$$

**4.3.4 Testing Results.** Results of classification accuracy are detailed in Table 7. Across all activities, the model achieves an accuracy of 84.3%. They were highest (100%) for washing the torso with a washcloth activity and lowest (66.7%) for the hammering a nail activity. We have also analyzed whether the parameters of the user’s activity (height, distance from the body, motion range, see Section 3.2) have an influence on how accurately an intended handover can be identified. The model has achieved the maximum accuracy for the activities performed close to the body (90.3%), and a somewhat lower accuracy for activities away from the body (79.5%). The accuracy of the model is also highest for activities comprising a large motion range (88.5%) and small motion range (85.0%), and slightly lower for activities with a medium motion range (79.2%). The height of activities does not affect classification accuracy ( $\approx 84\%$  for both head and torso levels).

Table 8 shows the result of the Joint Importance Rank (JIR) analysis. It reveals that the position of the left wrist and the rotation of the neck are the most impactful features in the primary user’s motion features that can convey that a handover is happening.

## 5 User Study: Validating the Perceived Quality of the Handover Interaction

We conducted a user study to compare the perceived quality of overall handover interactions generated by our data-driven method, trained on the 3HANDS dataset, with an established baseline method for performing handovers with an SRL [17]. To focus on validating the efficacy of our dataset and its capability to enable generative models while minimizing confounding variables potentially introduced by a specific hardware implementation, we carried out the user study in a virtual reality (VR) environment.

### 5.1 Experiment Design

The study employed a within-subject design. The participants were asked to perform handover interactions with a

Activity	Accuracy
Mount a mic	91.7%
Apply sunscreen to face	91.7%
Apply body lotion to chest	91.7%
Shampoo hair	83.3%
Wash torso with washcloth	100%
Blow dry hair	83.3%
Straighten a pic (low)	75.0%
Straighten a picture (high)	83.3%
Hammer a nail	66.7%
Clean a window	75.0%
Paint the wall (low)	81.3%
Paint the wall (high)	91.7%
Neutral pose	83.3%
<b>Overall</b>	<b>84.4%</b>

**Table 7: Classification accuracy for predicting the time frame of handover.**

JIR	Joint Position	Joint Rotation
1	Left wrist	Neck
2	Chest notch	Right shoulder
3	Head	Right elbow
4	Chest notch	Left elbow
5	Upper back	Face
6	Back	Back
7	Right hand	Left shoulder
8	Right elbow	Right wrist
9	Face	Left wrist
10	Right wrist	Right arm
11	Left arm	Left arm
12	Left elbow	Chest notch
13	Neck	Upper back
14	Right shoulder	Head
15	Left shoulder	Right hand
16	Right arm	Left hand

**Table 8: Joint Importance Rank (JIR) for predicting the timeframe of handover, indicating the relevance of an individual joint’s position and rotation.**

virtual SRL, where the SRL’s motions are generated by the baseline and 3HANDS data-driven methods. To constrain the overall study duration to an hour and still allow for two repetitions and a wide range of different motions, we selected 6 out of 13 total activities. The handover approaches were counterbalanced with a Balanced Latin Square to mitigate order effects. After experiencing one approach, the participants were asked to respond to eight 7-point Likert questions (see Figure 7), focusing on the aspects of naturalness, comfort, physical demand, predictability, timing, smoothness, and appropriateness. We recruited a total of 10 participants (5 male, 5 female, aged 16-58). The participants received monetary compensation for their participation in the study.

## 5.2 Motion Generation Approaches

We implemented the following two methods:

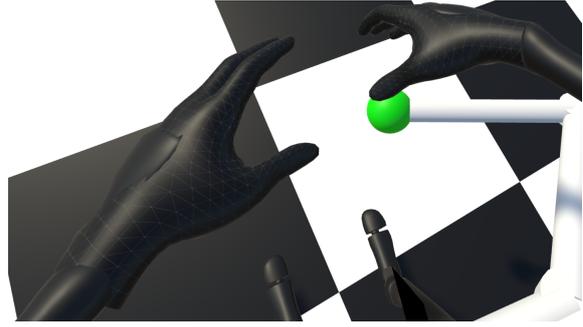
**5.2.1 Generative Models Trained with the 3HANDS Dataset.** Our approach integrates the **trajectory generating** SVAE model (subsection 4.1) and the **handover timeframe predicting** model (subsection 4.3). The **handover timeframe predicting** model continuously monitors the user's motions to determine whether the user is initiating a handover process or engaged in another activity. Once the **handover-timeframe-predicting** model detects that the user has initiated a handover process, the SVAE model then begins generating SRL's motion autoregressively, based on both the current and past motions of the user and the SRL. The handover is completed when the SRL's end-effector comes within close distance of the user's hand, at which point the SRL halts its motion to finalize the transfer. To prevent the object from entering the hand's simulation and causing object-hand collisions, we set the distance threshold to 12 cm, taking into account the object's size of 10 cm. In the following, we refer to our approach as *3HANDS*.

**5.2.2 Baseline Approach.** The most established approach to drive an SRL to complete handover motion is by predicting the user's hand position via extrapolation. We chose a baseline implementation from closely related prior work that shares the same SRL-centric setting with our scenario [17]. In this approach, pre-defined activation regions in the workspace serve as triggers for the SRL. When the user places their hand within one of these predefined 3D volumes, the SRL recognizes the user's intention to initiate a handover. Once activated, the SRL relies on a Kalman filter to predict the next 3D position of the user's hand until the handover is complete.

Using the predicted 3D position, the SRL calculates a trajectory to approach the user's hand. The SRL stops its movement when it reaches a predefined distance (12 cm in our implementation) from the user's hand, waiting for the object to be transferred. The SRL in the original paper utilized a 6-degree-of-freedom (6DoF) configuration, with 3DoF dedicated to reaching the goal position and the additional 3DoF used for collision avoidance. However, in our study, the SRL has a 3-degree-of-freedom (3DoF) configuration, focusing solely on the end-effector's position because collision avoidance is not the focus of our study. In the following, we refer to this approach as *baseline*.

## 5.3 Apparatus and Task

The experimental setup to evaluate the handover approaches was implemented in a virtual reality (VR) environment using Unity3D, run on a Quest Pro VR set. The whole experiment was run on Windows 10 with NVIDIA GeForce RTX 4090 GPU. The participants observed a humanoid representation of themselves in VR, and an SRL was virtually mounted on their hip (see Figure 6). The object for the handover was a sphere positioned at the end-effector of the SRL. During the experiment, the user's and SRL's current poses were transmitted to the model at each time step. The model then generated the SRL's subsequent position, both during handover interactions and while the SRL remained idle.



**Figure 6: First person view in VR of the handover interaction: Participants performed a task and then instructed the SRL to hand over the green ball either using the 3HANDS or baseline method.**

## 5.4 Procedure

Participants were first introduced to the study, followed by a tutorial for both approaches. Then, each participant performed 24 handover trials (2 approaches, 6 tasks, 2 repetitions). After each trial, they answered the 8 questions.

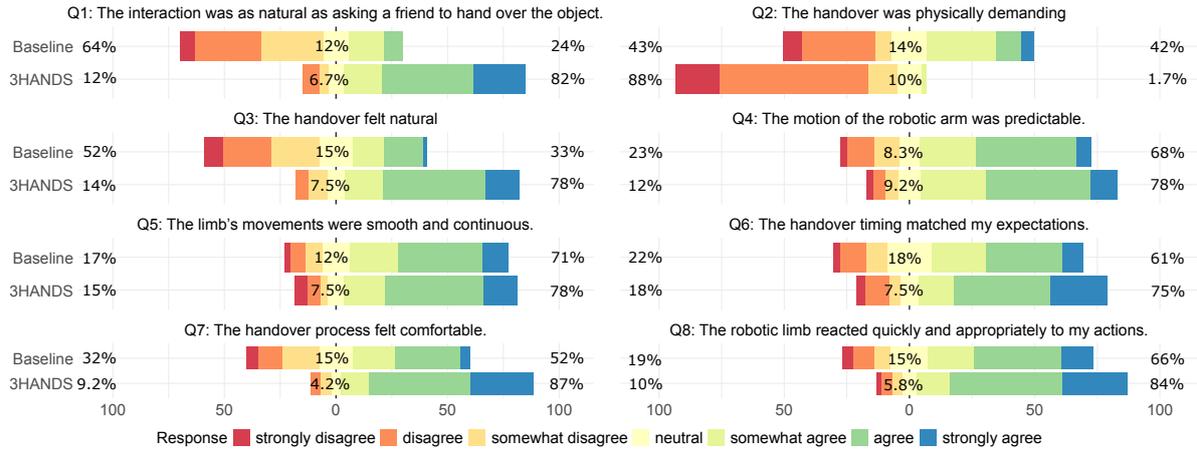
## 5.5 Results and Discussion

We analyzed the Likert ratings using Wilcoxon Signed Rank Tests. The results are presented in Figure 7.

**5.5.1 Perceived Naturalness.** Question 1 (*The interaction was as natural as asking a friend to hand over the object*) and question 3 (*The handover felt natural*) focus on the user's perceived naturalness. We asked two questions to capture two facets of naturalness: Q1 focuses on interpreting naturalness as interacting with a friendly human, while question 3 leaves more room for a broader interpretation. We found significant effects for both question 1 (3HANDS median = 6, baseline median = 3,  $W = 2086$ ,  $p < .001$ ) and question 3 (3HANDS median = 6, baseline median = 3,  $W = 3197$ ,  $p < .001$ ). The results indicated that 3HANDS allows for higher perceived naturalness.

**5.5.2 Perceived Physical Demand and Comfort.** Question 2 (*The handover was physically demanding*) and question 7 (*The handover process felt comfortable*) are related to the perceived physical demand and comfort. We found a significant difference in question 2 (3HANDS median = 2, baseline median = 4,  $W = 10855$ ,  $p < .001$ ), indicating 3HANDS allows for generative models for motions with less perceived demands. Similarly, we found a significant effect for question 7 (3HANDS median = 6, baseline median = 5,  $W = 3538.5$ ,  $p < .001$ ) which indicates that 3HANDS led to higher perceived comfort.

**5.5.3 Predictability and Smoothness.** Question 4 (*The motion of the robotic arm was predictable*) and question 5 (*The limb's movements were smooth and continuous*) examined the perceived predictability and motion smoothness. Wilcoxon Signed Rank Tests showed no significant differences between 3HANDS and baseline (both  $p > 0.05$ ). Medians of 5



**Figure 7: Results of the 7-point Likert scale for qualitative questions comparing the data-driven 3HANDS and baseline approaches.**

for baseline and 6 for 3HANDS indicate promising handover smoothness and predictability for both approaches.

**5.5.4 Perceived Timeliness and Appropriateness.** Finally, question 6 (*The handover timing matched my expectations*) and question 8 (*The robotic limb reacted quickly and appropriately to my actions*) validated the perceived timeliness and appropriateness of the handover motion. For question 6, a significant difference was found between 3HANDS and baseline (3HANDS median = 6, baseline median = 5,  $W = 5512.5$ ,  $p < .01$ ). We also found a significant difference in question 8 (3HANDS median = 6, baseline median = 5,  $W = 5170.5$ ,  $p < .001$ ). These findings highlight that the data-driven method is better in timing and more appropriate motion.

**5.5.5 Summary.** To conclude, our approach generated more natural, more comfortable, more timely, and more appropriate handover interaction motions compared to the baseline. The results highlight the application opportunities of our 3HANDS dataset and the presented generative models.

## 6 Discussion and Limitations

Our results demonstrate that the dataset effectively captures the key features of human-to-human, asymmetric, and asynchronous handover motions, making it well-suited for training SRLs. Notably, our dataset enables models to accurately generate handover motions, predict the handover region, and determine the timing of the handover event. The positive results in the user study indicated the models trained with the 3HANDS dataset generally result in more natural, more comfortable, and smoother handover interactions. Future research should explore more complex architectures.

An important future direction is to implement our model in practical applications. Future research should deploy our data-driven handover models in physical SRLs. One potential challenge is bridging the gap between synthetic environments and real-world conditions; exploring various techniques for improving simulation-to-reality (sim2real)

transfer will be required. Additionally, rapid and robust control methods coupled with accurate motion sensors may be essential to match the generated physical motions with the desired outcomes. Moreover, the development of safety-aware generative models is critical to ensure that predicted trajectories are compatible with safe and reliable robot operation.

Our dataset has further potential for applications beyond the models we explored in this paper. One promising avenue is to analyze human-to-human handover dynamics when primary activities are performed simultaneously with implicit handover requests. Research could shed new light on the intricacies of human communication and collaboration by examining how humans coordinate and complete tasks while handing objects to others. This understanding could be utilized in various contexts, such as developing more natural and intuitive interfaces for multi-user scenarios. Furthermore, our dataset could also serve as a basis for generating realistic human-to-avatar handovers in virtual environments. An exciting and timely direction would be the deployment of our models in virtual reality, where virtual arms could perform object handovers with human users. This application reduces the need for the precise control methods required in the physical world but presents the challenge of rendering realistic motions for objects of varying and unconstrained properties and from different directions. By incorporating the captured nuances in avatar interactions, we can create more immersive and engaging virtual environments that better reflect social dynamics. Increasing the predictability of object interactions from reliable real world data furthermore helps reduce the cognitive load during VR interactions in the absence of additional feedback modalities (like haptics) that are available in real-world SRL interactions. Addressing this challenge will be key to expanding the model's use in virtual environments. Additionally, our dataset may be leveraged to train the arm motion of mobile robots to exhibit human-like motion patterns around and near humans, enabling them to navigate complex social situations more easily and naturally.

While our study successfully demonstrates the feasibility of recording human motions for robotic trajectory generation, we acknowledge that the setup exhibits a downward bias in handover locations and human enactment may not always fully capture the nuances of robotic motion, particularly due to the differences in morphology and origin between another person's arm and a specific implementation of a body-worn robotic arm. Moreover, the social dynamics of two interacting humans may not be fully representative of interactions with a robotic arm, as the latter might be perceived as inherent extensions of one's own body. We attempted to mitigate this issue to the extent possible by only recording data from couples who live in a stable relationship and hence interact comfortably in close peripersonal space, and by shielding the robot participant's head from the primary user's view.

Future research should consider using our models as pre-trained baselines for fine-tuning in specific applications. For example, researchers could fine-tune our model for handover motions in different robotic arm settings (e.g., shoulder-mounted or environment-mounted).

Additionally, our dataset focuses on trajectories without including object-specific details. Different objects may require distinct handling strategies, affecting parameters such as grasp, object orientation, and motion speed. Future work should fine-tune the model with object properties so that it can generate motions tailored to the affordances of different objects, such as mugs containing liquid or heavy items. Because our models are trained on general handover tasks, they support transfer learning with minimal data. While incorporating object affordances and hand interactions are two possible extensions to the handover space defined by 3HANDS, their coverage is out of the main focus of this work. We acknowledge their potential for future work and have included hand joints tracking in the 3HANDS to support future exploration. Furthermore, our user study on perceived motion quality is deployed in a VR setting to ensure the effects of the confounding variables raised by physical environments are minimized. Yet, future research should integrate sophisticated control methods to bring such handover interactions from VR to the physical realm.

Moreover, our positive outcomes suggest that subsets of our dataset can be used to train lightweight models. For instance, researchers could focus on the most critical joints identified in our experiments (e.g., left hand, left wrist) and develop models accordingly. By gathering additional data on these key joints, transfer learning can be further applied to specific tasks. Finally, our joint importance analysis offers valuable insights for future data collection in other handover scenarios. In situations where full-body tracking is not feasible, this analysis can guide sensor placement decisions to optimize data collection and deployment of interactive systems.

Finally, a significant and novel challenge we identified is that the autoregressive generation of robot motions does not closely align with ground truth data. This challenge arises because the model generates the robot's motions while observing only the primary user's movements—an issue that significantly differs from existing motion generation tasks. Addressing this challenge may necessitate more complex

models. One potential solution is to enhance the model's reasoning capabilities, allowing it to better interpret the primary user's intentions (e.g., waiting to receive an object, moving to a destination, being occupied, etc.) and use this additional context to inform the motion generation. Another approach could be the integration of reinforcement learning, which could train a policy model to adapt to the environment and the primary user.

## 7 Conclusion

In this paper, we have presented the 3HANDS dataset, which provides extensive capture of object handovers between closely interacting humans. It considerably extends beyond prior datasets by its asymmetric spatial configuration with handovers occurring in intimate peripersonal space, the participant's asymmetric roles, real-world primary activities, and implicit coordination of handover. This is representative of the unique demands of handovers between humans and wearable robotic limbs. 12 unique pairings of participants were captured in 41 synchronized 2K camera views, from which we calculate rigged 3D skeleton data and hand poses. The dataset also includes transcripts of utterances, such as verbal commands and reactions, as well as manually annotated ground-truth data for object handover.

In a series of experiments, we demonstrate the applicability of the dataset for training models for interaction with SRLs. We contribute models and corresponding technical evaluation results that each address one key aspect of a handover activity. We contribute a generative model, based on a conditional variational autoencoder, which generates the trajectory of a handover in response to the primary user's motion. Furthermore, we present a model that can accurately generate the region of transfer, where an object will be handed over. Additionally, we show that using our dataset it is possible to accurately predict, solely from implicit user posture, when the handover should be initiated. Finally, we deployed our models for performing handover interactions compared to an established baseline approach in a VR setting. The user study showed that our data-driven approach enables more natural and comfortable handover interaction, further highlighting the potential value of 3HANDS for training SRL models.

We share the dataset with the community to foster future research on interactive systems and to help deepen the understanding of the unique characteristics of handover activities in close personal space.

## References

- [1] Mohammed Al-Sada, Thomas Höglund, Mohamed Khamis, Jaryd Urbani, and Tatsuo Nakajima. 2019. Orochi: Investigating Requirements and Expectations for Multipurpose Daily Used Supernumerary Robotic Limbs. In *Proceedings of the 10th Augmented Human International Conference 2019 (AHS '19)*. Association for Computing Machinery, New York, NY, USA, Article 37, 9 pages. <https://doi.org/10.1145/3311823.3311850>
- [2] Mohammed Al-Sada, Keren Jiang, Shubhankar Ranade, Mohammed Kalkattawi, and Tatsuo Nakajima. 2020. HapticSnakes: Multi-haptic Feedback Wearable Robots for Immersive Virtual Reality. *Virtual Reality* 24, 2 (01 Jun 2020), 191–209. <https://doi.org/10.1007/s10055-019-00404-x>
- [3] Ali Al-Yacoub, Myles Flanagan, Achim Buerkle, Thomas Bamber, Pedro Ferreira, Ella-Mae Hubbard, and Niels Lohse. 2021. Data-driven modelling of human-human co-manipulation using force and muscle surface electromyogram activities. *Electronics* 10, 13 (2021), 1509.

- [4] Jacopo Aleotti, Vincenzo Micelli, and Stefano Caselli. 2014. An affordance sensitive system for robot to human object handover. *International Journal of Social Robotics* 6 (2014), 653–666.
- [5] Ken Arai, Hiroto Saito, Masaaki Fukuoka, Sachiyo Ueda, Maki Sugimoto, Michiteru Kitazaki, and Masahiko Inami. 2022. Embodiment of Supernumerary Robotic Limbs in Virtual Reality. *Scientific Reports* 12, 1 (27 Jun 2022), 9769. <https://doi.org/10.1038/s41598-022-13981-w>
- [6] Michael Argyle and Janet Dean. 1965. Eye-contact, distance and affiliation. *Sociometry* (1965), 289–304.
- [7] Patrizia Basili, Markus Huber, Thomas Brandt, Sandra Hirche, and Stefan Glasauer. 2009. Investigating human-human approach and handover. *Human centered robot systems: Cognition, interaction, technology* (2009), 151–160.
- [8] Alessandro Carfi, Francesco Fogliano, Barbara Bruno, and Fulvio Mastrogiovanni. 2019. A multi-sensor dataset of human-human handover. *Data in brief* 22 (2019), 109–117.
- [9] Wesley P Chan, Matthew KXJ Pan, Elizabeth A Croft, and Masayuki Inaba. 2020. An affordance and distance minimization based method for computing object orientations for robot human handovers. *International Journal of Social Robotics* 12, 1 (2020), 143–162.
- [10] Wesley P Chan, Matthew KXJ Pan, Elizabeth A Croft, and Masayuki Inaba. 2020. An affordance and distance minimization based method for computing object orientations for robot human handovers. *International Journal of Social Robotics* 12, 1 (2020), 143–162.
- [11] Wesley P. Chan, Tin Tran, Sara Sheikholeslami, and Elizabeth Croft. 2021. An Experimental Validation and Comparison of Reaching Motion Models for Unconstrained Handovers: Towards Generating Humanlike Motions for Human-Robot Handovers. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. 356–361. <https://doi.org/10.1109/HUMANOIDS47582.2021.9555779>
- [12] S.H. Cheong, J.H. Lee, and C.H. Kim. 2018. A New Concept of Safety Affordance Map for Robots Object Manipulation. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 565–570. <https://doi.org/10.1109/ROMAN.2018.8525627>
- [13] Francesca Cini, V Orteni, P Corke, and MJSR Controzzi. 2019. On the choice of grasp type and location when handing over an object. *Science Robotics* 4, 27 (2019), eaa9757.
- [14] Markos Diomatari, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J Black. 2024. WANDR: Intention-guided Human Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 927–936.
- [15] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.
- [16] Yuwei Du, Xin Zhang, Mattia Leonori, Pietro Balatti, Jing Jin, Qiang Wang, and Arash Ajoudani. 2023. Bi-Directional Human-Robot Handover Using a Novel Supernumerary Robotic System. In *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. 153–158. <https://doi.org/10.1109/ARSO56563.2023.10187506>
- [17] Yuwei Du, Xin Zhang, Mattia Leonori, Pietro Balatti, Jing Jin, Qiang Wang, and Arash Ajoudani. 2023. Bi-Directional Human-Robot Handover Using a Novel Supernumerary Robotic System. In *2023 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. 153–158. <https://doi.org/10.1109/ARSO56563.2023.10187506>
- [18] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. 2023. IMoS: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, 1–12.
- [19] Michael J Gielniak, C Karen Liu, and Andrea L Thomaz. 2013. Generating human-like motion for robots. *The International journal of robotics research* 32, 11 (2013), 1275–1301.
- [20] Stefan Glasauer, Markus Huber, Patrizia Basili, Alois Knoll, and Thomas Brandt. 2010. Interacting in time and space: Investigating human-human and human-robot joint action. In *19th international symposium in robot and human interactive communication*. IEEE, 252–257.
- [21] Edmund T Hall and Edward T Hall. 1966. *The hidden dimension*. Vol. 609. Anchor.
- [22] Clint Hansen, Paula Arambel, Khalil Ben Mansour, Véronique Perdereau, and Frédéric Marin. 2017. Human–human handover tasks and how distance and object mass matter. *Perceptual and motor skills* 124, 1 (2017), 182–199.
- [23] Clint Hansen, Paula Arambel, Khalil Ben Mansour, Véronique Perdereau, and Frédéric Marin. 2017. Human–human handover tasks and how distance and object mass matter. *Perceptual and motor skills* 124, 1 (2017), 182–199.
- [24] Steen Harsted, Greg Kawchuk, Raymond Guan, Tue Skallgård, Bue Bonderup Hesby, Eleanor Boyle, and Per Kjaer. 2019. The performance of two in-clinic markerless motion capture systems compared to a laboratory standard. In *WFC Biannual Congress/ECU Convention 2019*.
- [25] Neville Hogan. 1982. Control and Coordination of Voluntary Arm Movements. In *1982 American Control Conference*. 522–528. <https://doi.org/10.23919/ACC.1982.4787906>
- [26] Chien-Ming Huang, Maya Cakmak, and Bilge Mutlu. 2015. Adaptive Coordination Strategies for Human-Robot Handovers.. In *Robotics: science and systems*, Vol. 11. Rome, Italy, 1–10.
- [27] Chien-Ming Huang, Maya Cakmak, and Bilge Mutlu. 2015. Adaptive Coordination Strategies for Human-Robot Handovers.. In *Robotics: science and systems*, Vol. 11. Rome, Italy, 1–10.
- [28] Markus Huber, Markus Rickert, Alois Knoll, Thomas Brandt, and Stefan Glasauer. 2008. Human-robot interaction in handing-over tasks. In *RO-MAN 2008-the 17th IEEE international symposium on robot and human interactive communication*. IEEE, 107–112.
- [29] Irfan Hussain, Gionata Salvietti, Giovanni Spagnoletti, and Domenico Prattichizzo. 2016. The soft-sixthfinger: a wearable emg controlled robotic extra-finger for grasp compensation in chronic stroke patients. *IEEE Robotics and Automation Letters* 1, 2 (2016), 1000–1006.
- [30] Matthew S Johannes, John D Bigelow, James M Burck, Stuart D Harshbarger, Matthew V Kozlowski, and Thomas Van Doren. 2011. An Overview of the Developmental Process for the Modular Prosthetic Limb. *Johns Hopkins APL Technical Digest* 30, 3 (2011), 207–216.
- [31] Tadakazu Kashiwabara, Hirotaka Osawa, Kazuhiko Shinozawa, and Michita Imai. 2012. TEROOS: A Wearable Avatar to Enhance Joint Activities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) (CHI '12). Association for Computing Machinery, New York, NY, USA, 2001–2004. <https://doi.org/10.1145/2207676.2208345>
- [32] Parag Khanna, Márten Björkman, and Christian Smith. 2022. Human inspired grip-release technique for robot-human handovers. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 694–701.
- [33] Parag Khanna, Márten Björkman, and Christian Smith. 2023. A multi-modal data set of human handovers with design implications for human-robot handovers. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1843–1850.
- [34] Ingrid Kovacs, Alexandra Iosub, Marina Topa, Andi Buzo, and Georg Pelz. 2019. A gradient-based sensitivity analysis method for complex systems. In *2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME)*. IEEE, 333–338.
- [35] Robert Kovacs, Eyal Ofek, Mar Gonzalez Franco, Alexa Fay Siu, Sebastian Marwecki, Christian Holz, and Mike Sinclair. 2020. Haptic PIVOT: On-Demand Handhelds in VR. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) (UIST '20). Association for Computing Machinery, New York, NY, USA, 1046–1059. <https://doi.org/10.1145/3379337.3415854>
- [36] Alap Kshirsagar, Raphael Fortuna, Zhiming Xie, and Guy Hoffman. 2023. Dataset of bimanual human-to-human object handovers. *Data in Brief* 48 (2023), 109277.
- [37] Alap Kshirsagar, Hadas Kress-Gazit, and Guy Hoffman. 2019. Specifying and synthesizing human-robot handovers. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5930–5936.
- [38] D. Lambert and Diagram Group. 2004. *Body Language*. HarperCollins. <https://books.google.de/books?id=jo0EAAAACAAJ>
- [39] Chiara Talignani Landi, Yujiao Cheng, Federica Ferraguti, Marcello Bonfè, Cristian Secchi, and Masayoshi Tomizuka. 2019. Prediction of Human Arm Target for Robot Reaching Movements. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5950–5957. <https://doi.org/10.1109/IROS40897.2019.8968559>
- [40] Jungpyo Lee, Licheng Yu, Lucie Derbier, and Hannah S. Stuart. 2021. Assistive Supernumerary Grasping with the Back of the Hand. In *2021 IEEE International Conference on Robotics and Automation (ICRA '21)*. IEEE, 6154–6160. <https://doi.org/10.1109/ICRA48506.2021.9560949>
- [41] Benedikt Leichtmann and Verena Nitsch. 2020. How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *Journal of Environmental Psychology* 68 (2020), 101386. <https://doi.org/10.1016/j.jenvp.2019.101386>
- [42] Benedikt Leichtmann and Verena Nitsch. 2020. How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *Journal of environmental Psychology* 68 (2020), 101386.
- [43] Sang-won Leigh, Timothy Denton, Kush Parekh, William Peebles, Magnus Johnson, and Pattie Maes. 2018. Morphology Extension Kit: A Modular Robotic Platform for Physically Reconfigurable Wearables. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (Stockholm, Sweden) (TEI '18). Association for Computing Machinery, New York, NY, USA, 11–18. <https://doi.org/10.1145/3173225.3173239>
- [44] Sang-won Leigh and Pattie Maes. 2016. Body Integrated Programmable Joints Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 6053–6057. <https://doi.org/10.1145/2858036.2858538>

- [45] Zhi Li and Kris K. Hauser. 2015. Predicting Object Transfer Position and Timing in Human-robot Handover Tasks. <https://api.semanticscholar.org/CorpusID:53649936>
- [46] Xuwei Lin, Xiaohui Xiao, and Zhao Guo. 2021. Mechanical Design of a Supernumerary Robotic Finger for Grasping Abilities Compensation. In *2021 IEEE International Conference on Robotics and Biomimetics (RO-BIO '21)*. IEEE, 1792–1797. <https://doi.org/10.1109/ROBIO54168.2021.9739568>
- [47] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. 2020. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 40–1.
- [48] Baldin Llorens-Bonilla and H. Harry Asada. 2014. A Robot on the Shoulder: Coordinated Human-wearable Robot Control Using Coloured Petri Nets and Partial Least Squares Predictions. *2014 IEEE International Conference on Robotics and Automation (2014)*, 119–125.
- [49] Baldin Llorens-Bonilla, Federico Parrietti, and H Harry Asada. 2012. Demonstration-based Control of Supernumerary Robotic Limbs. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '12)*. IEEE, 3936–3942. <https://doi.org/10.1109/IROS.2012.6386055>
- [50] Jianwen Luo, Zelin Gong, Yao Su, Lecheng Ruan, Ye Zhao, H. Harry Asada, and Chenglong Fu. 2021. Modeling and Balance Control of Supernumerary Robotic Limb for Overhead Tasks. *IEEE Robotics and Automation Letters* 6, 2 (2021), 4125–4132. <https://doi.org/10.1109/LRA.2021.3067850>
- [51] Jianwen Luo, Sicong Liu, Chengyu Lin, Yong Zhou, Zixuan Fan, Zheng Wang, Chaoyang Song, H Harry Asada, and Chenglong Fu. 2021. Mapping Human Muscle Force to Supernumerary Robotics Device for Overhead Task Assistance. *arXiv preprint arXiv:2107.13799* (2021).
- [52] Guilherme Maeda, Marco Ewerton, Gerhard Neumann, Rudolf Lioutikov, and Jan Peters. 2017. Phase estimation for fast action recognition and trajectory generation in human-robot collaboration. *The International Journal of Robotics Research* 36, 13-14 (2017), 1579–1594. <https://doi.org/10.1177/0278364917693927>
- [53] Andrea H Mason and Christine L MacKenzie. 2005. Grip forces when passing an object to a partner. *Experimental brain research* 163 (2005), 173–187.
- [54] José R. Medina, Felix Duvallet, Murali Karnam, and Aude Billard. 2016. A human-inspired controller for fluid human-robot handovers. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 324–331. <https://doi.org/10.1109/HUMANOIDS.2016.7803296>
- [55] Thomas Most and Johannes Will. 2024. Sensitivity analysis using the Metamodel of Optimal Prognosis. *arXiv preprint arXiv:2408.03590* (2024).
- [56] Marie Muehlhaus, Marion Koelle, Artin Saberpour, and Jürgen Steimle. 2023. I Need a Third Arm! Eliciting Body-based Interactions with a Wearable Robotic Arm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3544548.3581184>
- [57] Jonathan Mumm and Bilge Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction (Lausanne, Switzerland) (HRI '11)*. Association for Computing Machinery, New York, NY, USA, 331–338. <https://doi.org/10.1145/1957656.1957786>
- [58] Sina Parastegari, Bahareh Abbasi, Ehsan Noohi, and Miloš Zefran. 2017. Modeling human reaching phase in human-human object handover with application in robot-human handover. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3597–3602.
- [59] Federico Parrietti, Kameron C Chan, Banks Hunter, and H Harry Asada. 2015. Design and control of supernumerary robotic limbs for balance augmentation. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5010–5017.
- [60] Vignesh Prasad, Dorothea Koert, Ruth Stock-Homburg, Jan Peters, and Georgia Chalvatzaki. 2022. MILD: Multimodal Interactive Latent Dynamics for Learning Human-Robot Interaction. In *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. 472–479. <https://doi.org/10.1109/Humanoids53995.2022.10000239>
- [61] Vignesh Prasad, Alap Kshirsagar, Dorothea Koert, Ruth Stock-Homburg, Jan Peters, and Georgia Chalvatzaki. 2024. MoVEInt: Mixture of Variational Experts for Learning Human-Robot Interactions From Demonstrations. *IEEE Robotics and Automation Letters* 9, 7 (2024), 6043–6050. <https://doi.org/10.1109/LRA.2024.3396074>
- [62] Domenico Prattichizzo, Maria Pozzi, Tommaso Lisini Baldi, Monica Malvezzi, Irfan Hussain, Simone Rossi, and Gionata Salvietti. 2021. Human augmentation by wearable supernumerary robotic limbs: review and perspectives. *Progress in Biomedical Engineering* 3, 4 (2021), 042005.
- [63] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. 2021. Humor: 3d human motion model for robot pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11488–11499.
- [64] Artin Saberpour Abadian, Ata Otaran, Martin Schmitz, Marie Muehlhaus, Rishabh Dabral, Diogo Luvizon, Azumi Maekawa, Masahiko Inami, Christian Theobalt, and Jürgen Steimle. 2023. Computational Design of Personalized Wearable Robotic Limbs. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>)* (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 68, 13 pages. <https://doi.org/10.1145/3586183.3606748>
- [65] MHD Yamen Sarajji, Tomoya Sasaki, Kai Kunze, Kouta Minamizawa, and Masahiko Inami. 2018. MetaArms: Body Remapping Using Feet-Controlled Artificial Arms. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (Berlin, Germany) (UIST '18)*. Association for Computing Machinery, New York, NY, USA, 65–74. <https://doi.org/10.1145/3242587.3242665>
- [66] Tomoya Sasaki, MHD Yamen Sarajji, Charith Lasantha Fernando, Kouta Minamizawa, and Masahiko Inami. 2017. MetaLimbs: Multiple Arms Interaction Metamorphism. In *ACM SIGGRAPH 2017 Emerging Technologies*. 1–2.
- [67] Leonardo Sabatino Scimmi, Matteo Melchiorre, Stefano Mauro, and Stefano Pastorelli. 2019. Experimental real-time setup for vision driven hand-over with a collaborative robot. In *2019 International Conference on Control, Automation and Diagnosis (ICCAD)*. IEEE, 1–5.
- [68] Sara Sheikholeslami, Gilwoo Lee, Justin W. Hart, Siddhartha Srinivasa, and Elizabeth A. Croft. 2020. A Study of Reaching Motions for Collaborative Human-Robot Interaction. In *Proceedings of the 2018 International Symposium on Experimental Robotics*, Jing Xiao, Torsten Kröger, and Oussama Khatib (Eds.). Springer International Publishing, Cham, 584–594.
- [69] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).
- [70] Roy Someshwar and Yael Edan. 2017. Givers & Receivers perceive handover tasks differently: Implications for Human-Robot collaborative system design. *arXiv preprint arXiv:1708.06207* (2017).
- [71] Sebastian Starke, Ian Mason, and Taku Komura. 2022. Deepphase: Periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.
- [72] Carsten Stoll, Nils Hasler, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. 2011. Fast articulated motion tracking using a sums of Gaussians body model. In *2011 International Conference on Computer Vision*.
- [73] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, and Siddhartha S Srinivasa. 2012. Learning the communication of intent prior to physical collaboration. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 968–973.
- [74] Kyle Strabala, Min Kyung Lee, Anca Dragan, Jodi Forlizzi, Siddhartha S Srinivasa, Maya Cakmak, and Vincenzo Micelli. 2013. Toward seamless human-robot handovers. *Journal of Human-Robot Interaction* 2, 1 (2013), 112–132.
- [75] Leila Takayama and Caroline Pantofaru. 2009. Influences on proxemic behaviors in human-robot interaction. In *2009 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 5495–5502.
- [76] Yuchuang Tong and Jinguo Liu. 2021. Review of research and development of supernumerary robotic limbs. *IEEE/CAA Journal of Automatica Sinica* 8, 5 (2021), 929–952.
- [77] Vighnesh Vatsal and Guy Hoffman. 2017. Wearing your arm on your sleeve: Studying usage contexts for a wearable robotic forearm. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 974–980.
- [78] Vighnesh Vatsal and Guy Hoffman. 2018. Design and analysis of a wearable robotic forearm. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5489–5496.
- [79] Vighnesh Vatsal and Guy Hoffman. 2018. Design and Analysis of a Wearable Robotic Forearm. In *2018 IEEE International Conference on Robotics and Automation (ICRA '18)*. IEEE, 5489–5496. <https://doi.org/10.1109/ICRA.2018.8461212>
- [80] Catherine Véronneau, Jeff Denis, Louis-Philippe Lebel, Marc Denninger, Vincent Blanchard, Alexandre Girard, and Jean-Sébastien Plante. 2020. Multifunctional Remotely Actuated 3-DOF Supernumerary Robotic Arm Based on Magnetorheological Clutches and Hydrostatic Transmission Lines. *IEEE Robotics and Automation Letters* 5, 2 (2020), 2546–2553. <https://doi.org/10.1109/LRA.2020.2967327>
- [81] M.L. Walters, K. Dautenhahn, R. te Boekhorst, Kheng Lee Koay, C. Kaouri, S. Woods, C. Nehaniv, D. Lee, and I. Werry. 2005. The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005*. 347–352. <https://doi.org/10.1109/ROMAN.2005.1513803>

- [82] Weitian Wang, Rui Li, Yi Chen, Yi Sun, and Yunyi Jia. 2022. Predicting Human Intentions in Human–Robot Hand-Over Tasks Through Multimodal Learning. *IEEE Transactions on Automation Science and Engineering* 19, 3 (2022), 2339–2353. <https://doi.org/10.1109/TASE.2021.3074873>
- [83] Noah Wiederhold, Ava Megyeri, DiMaggio Paris, Sean Banerjee, and Natasha Banerjee. 2024. Hoh: Markerless multimodal human-object-human handover dataset with large object count. *Advances in Neural Information Processing Systems* 36 (2024).
- [84] Wouter Wolf, Jacques Launay, and Robin IM Dunbar. 2016. Joint attention, shared goals, and social bonding. *British Journal of Psychology* 107, 2 (2016), 322–337.
- [85] Katsu Yamane, Marcel Revfi, and Tamim Asfour. 2013. Synthesizing object receiving motions of humanoid robots with human motion database. In *2013 IEEE International Conference on Robotics and Automation*. IEEE, 1629–1636.
- [86] Bo Yang, Jian Huang, Xinxing Chen, Caihua Xiong, and Yasuhisa Hasegawa. 2021. Supernumerary Robotic Limbs: A Review and Future Outlook. *IEEE Transactions on Medical Robotics and Bionics* 3, 3 (2021), 623–639. <https://doi.org/10.1109/TMRB.2021.3086016>
- [87] CJ Yang, Jiafan Zhang, Ying Chen, YM Dong, and Y Zhang. 2008. A Review of Exoskeleton-type Systems and Their Key Technologies. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 222, 8 (2008), 1599–1612.
- [88] Wei Yang, Balakumar Sundaralingam, Chris Paxton, Ireteayo Akinola, Yu-Wei Chao, Maya Cakmak, and Dieter Fox. 2022. Model Predictive Control for Fluid Human-to-Robot Handovers. *2022 International Conference on Robotics and Automation (ICRA) (2022)*, 6956–6962. <https://api.semanticscholar.org/CorpusID:247922355>
- [89] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. 2021. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15762–15771.
- [90] Shigeo Yoshida, Tomoya Sasaki, Zenda Kashino, and Masahiko Inami. 2023. TOMURA: A Mountable Hand-Shaped Interface for Versatile Interactions. In *Proceedings of the Augmented Humans International Conference 2023 (Glasgow, United Kingdom) (AHs '23)*. Association for Computing Machinery, New York, NY, USA, 243–254. <https://doi.org/10.1145/3582700.3582719>
- [91] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5745–5753.